中国科学: 信息科学 2024 年 第 54 卷 第 5 期: 1267–1282
SCIENTIA SINICA Informationis
《中国科学》杂志社
SCIENCE CHINA PRESS
6G 专项研究进展专刊 · 论文

面向 6G 网络的高可靠低延时通信计算与控制

牛志升

清华大学电子工程系, 北京 100084
E-mail: niuzhs@tsinghua.edu.cn

收稿日期: 2023–11–01; 修回日期: 2024–02–21; 接受日期: 2024–03–08; 网络出版日期: 2024–04–30

国家重点研发计划 (批准号: 2020YFB1806605) 资助项目

摘要 未来 6G 网络的业务主体将转变为赋能千行百业的智能物联网应用, 因此在提供高可靠低延时通信 (ultra-reliable and low-latency communications, uRLLComm) 的同时, 还要实现高可靠低延时的智能计算 (uRLLComp) 与智能控制 (uRLLCtrl) (以下简称 uRLLC^3). 这对于传输环境、业务需求与网络资源配置均高度动态变化的移动网络来讲是一个巨大挑战. 本文面向 6G 时代的智慧网联与网联智能, 从排队论中几个超越直觉的定理出发, 尝试为 uRLLC^3 的实现提供一些有益启示, 并系统地给出实现高可靠低延时通信、计算与控制的理论与方法. 同时, 以未来智慧网联汽车的智能控制为例, 在描述信息新鲜度的信息年龄 (age-of-information, AoI) 性能指标的基础上, 提出一种新的信息时效性表征指标 —— 信息紧迫度 (urgency-of-information, UoI), 并在此基础上给出几种 UoI 意义上最优的移动终端调度与控制算法, 验证所提 UoI 指标的有效性.

关键词 移动通信, 6G, 高可靠低延时, 车联网, 信息年龄, 信息时效性

1 引言

未来 5G 和 6G 网络所面对的业务主体不再是人和人之间的通信, 而是诸如智能交通、智慧工厂、智能物流、智慧医疗、智慧海洋等千行百业的垂直应用 (verticals 或称 mission-critical 应用). 因此, 未来移动通信与网络除了要提供 Gb/s~Tb/s 级的 eMBB (enhanced mobile broadband) 业务以及每平方公里 100~1000 万连接的 mMTC (massive machine-type communication) 业务之外, 更重要的是需要提供 99.999%~99.99999% 传输可靠性以及低至 1 ms~100 μs 延时的 uRLLC (ultra-reliable and low-latency communication) 业务, 这对业务需求与传输环境均高度动态变化的移动通信网络来讲是一个全新的挑战. 虽然, 3GPP (3rd Generation Partnership Project) 在 R15 至 R18 版本中给出了一些解决方案, 但距离实现上述目标还有相当的距离.

引用格式: 牛志升. 面向 6G 网络的高可靠低延时通信计算与控制. 中国科学: 信息科学, 2024, 54: 1267–1282, doi: 10.1360/SSI-2023-0336
Niu Z S. uRLLC^3: ultra-reliable and low-latency communication, computing, and control for 6G networks (in Chinese). Sci Sin Inform, 2024, 54: 1267–1282, doi: 10.1360/SSI-2023-0336

与此同时, 垂直应用中将包含大量需要对终端进行远程监测与控制的应用, 如自动驾驶、工业物联网等, 它们需要实时感知或监控终端和网络的状态 (如临近车辆的位置、速度、加速度等信息), 并将最新的状态信息 (status information) 汇聚至控制中心 (信宿). 5G/6G 网络中的大规模天线 (massive MIMO) 系统也需要对信道状态信息 (channel state information, CSI) 进行实时估计, 并回传至基站. 因此, 除了通信过程的可靠性和延时之外, 还需要综合考虑感知 (sensing)、计算 (computing) 和更新 (updating) 过程的可靠性与延时. 与传统网络中用户端产生的信息 (user information) 不同, 这些状态信息在不断地随时空发生变化, 而且一般具有马尔可夫 (Markov) 性, 即状态发生变化后, 新的状态信息将可以完全取代旧的状态信息, 因此在信宿处只需要保留最新的状态信息即可, 而且越新鲜越好. 同时, 还需要对不同信息在不同情境下的重要程度作出评估, 使得可在网络资源受限的情况下优先传输重要程度高的信息, 因此, 网络中的传统性能指标, 如吞吐量、延时和丢包率等, 均不足以描述状态更新业务的需求. 如何针对状态更新业务的特点优化移动网络的资源配置和调度策略, 是未来移动通信网络面临的新挑战.

更进一步地, 面向未来大数据和智能信息处理能力无处不在的机器学习 (machine learning, ML) 与人工智能 (artificial intelligence, AI) 时代, 未来信息网络将实现从简单的互联互通向智慧网联 (smart networking) 的转变, 即高度智能化的业务终端和网络节点将全方位地参与到信息的转发和处理过程中 (in-network computing, 在网计算), 实现更高可靠性、更低延时和更高效能的信息交互. 为此, 各网络节点需要大量引入计算资源, 在进行分布式学习与决策的同时分流周边业务终端的计算任务, 形成一个通信与计算深度融合的网络. 可见, “高可靠低延时” 需要进一步延伸至计算环节, 且两者是紧密耦合在一起的, 即通信过程中伴随着计算, 计算过程中也伴随着通信, 为该问题的求解带来了更大的挑战.

综上所述, 高可靠低延时业务中的 “可靠性” 与 “延时” 均有着丰富的内涵, 也延伸至了计算与控制环节的可靠性和延时, 将通信、计算和控制过程紧密地耦合在了一起, 在要求 99.999%∼99.99999% 传输可靠性的同时, 还要保障低至 1 ms∼100 μs 的传输延时以及近乎为零的延时抖动 (即所谓的 “确定性时延”), 这对传输环境高度动态变化的移动通信来讲是一个巨大挑战, 因此仅靠无线传输层技术的提高难以实现, 需要联合网络层和业务层技术, 同时借助算力网络与人工智能算法等共同解决. 为表述方便, 以下简称为 uRLLC^3 (ultra-reliable and low-latency communication, computing, and control).

本文首先对可靠性与延时的内涵进行扩展, 分别给出通信过程不同层次的可靠性与延时定义. 然后, 将 uRLLC 概念拓展到计算与控制环节, 并以未来智慧网联汽车的智能控制为例, 在描述信息新鲜度的信息年龄 (age-of-information, AoI) 性能指标的基础上, 提出一种新的时效性 (timeliness) 表征指标 —— 信息紧迫度 (urgency-of-information, UoI), 并在此基础上给出几种 UoI 意义上最优的移动终端调度与控制算法, 验证所提 UoI 指标的有效性.

2 高可靠性低延时: 内涵扩展

随着 5G, 6G 网络的业务主体从传统的人与人通信逐渐过渡到万物互联的垂直应用, 网络中传递的信息不再只是对网络透明的终端用户信息 (user information), 更多的将是为了实现某个垂直应用而需要交互的用户及网络状态信息 (status information) 以及在网计算 (in-network computing) 过程中需要交互的数据或模型信息 (data/model information) 等, 因此可靠性与延时的性能要求应同时涵盖通信、计算与控制过程. 一个典型的垂直应用是如图 1 所示的需要由多个智能体联合完成的协同感知 (multi-agent cooperative perception) 任务, 在此过程中, 每个智能体 (信源) 均会对感知对象从不同的

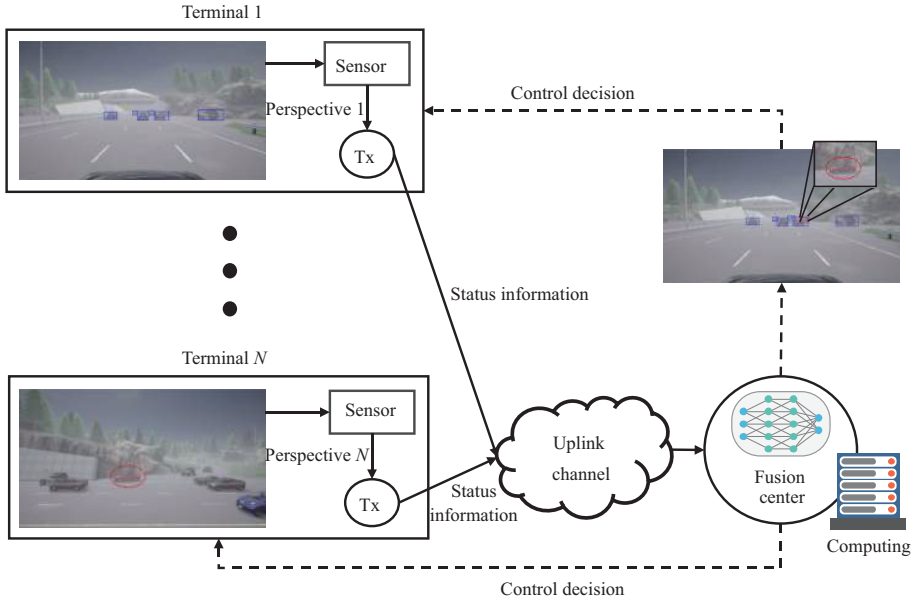

图 1 (网络版彩图) 多智能体协同环境感知系统示意图

Figure 1 (Color online) Schematic diagram of the multi-agent cooperative environment perception system

视角进行感知, 然后将感知到的信息汇聚到一个汇聚中心 (fusion center). 汇聚中心 (信宿) 则会通过机器学习等人工智能方法进行训练 (training) 和推断 (inference), 并将感知结果反馈给需求方. 如果感知精度未达到要求的精度或是感知对象发生了变化, 汇聚中心则会通知智能体重新上报最新的感知结果, 并不断迭代, 直至达到满意的感知精度.

可见, uRLLC^3 业务中的 “可靠性” 与 “延时” 均有着丰富的内涵, 远远超出了通信过程的可靠性和延时, 而且两者是相互耦合在一起的, 需要更细致地进行分解.

首先, 可靠性 (reliability) 指的是在规定的时间内成功传输的概率, 它可以依次被细分成 4 个层次.

Level-1: 随时随地可以传输, 即任何时间任何地点均有网络覆盖 (connectivity).

Level-2: 每次传输均能成功, 即尽量无须重传, 保障一定的吞吐量 (throughput).

Level-3: 每次传输均能在给定时间内成功传输, 即保障给定时间内的有效吞吐量 (timely throughput).

Level-4: 每次传输均能在给定时间内成功且确定性地传输, 即同时保障有效吞吐量和确定性延时 (timely and deterministic throughput). 这实际上相当于保障传输延时的可预测性, 需要在提高传输成功率的同时降低延时抖动 (jitter).

相应地, 延时 (latency) 指的是将信息尽可能快地发送到接收端, 它也可以细分为 4 个层次.

Level-1: 空口传输延时 (air-interface delay), 即仅包含接入侧的空中传输延时.

Level-2: 发端到收端的通信延时 (communication delay), 即通过整个通信网络的传输延时.

Level-3: 信息生成到信息使用的信息延时 (information delay), 即信息从被感知或生成时刻到被信息使用方正确接收为止的延时. 在万物智联、协同感知和远程控制等系统中, 该延时反映了汇聚中心或是控制中心 (信息使用方) 所掌握的状态信息与当前所关注终端 (信息生成方) 真实状态信息之间的滞后 (lapse) 程度或称时间误差 (temporal distortion) 程度, 它包含了感知、调度、传输等环节的延时.

Level-4: 信息生成到信息使用的信息时效性 (information timeliness), 即信息使用方当前所保有状

态信息的时间效用, 它既关注信息延时本身, 又关注不同大小和不同情境下的信息延时对系统性能的影响程度. 如果将信息延时视为状态信息的时间误差, 则信息时效性反映的是状态信息时间误差对系统服务效用的影响, 且不同的误差大小在不同的情境下对该效用的影响可能存在很大的差别.

以智能自主驾驶应用为例, 车辆 (信息生成方) 在前进过程中向路边单元上报实时位置、移动方向、移动速度、加速度等状态信息, 路边单元 (信息使用方) 根据其自身以及周边车辆的状态信息通过优化或人工智能等计算方法进行控制决策, 并将决策结果发送给车辆以控制车辆按照预定轨迹移动. 由于车辆数目众多、无线传输资源受限, 路边单元需要通过调度的方式控制车辆上报信息的顺序, 使得车辆的状态信息难以在发生变化时及时得到更新, 因此路边单元的决策也难以保障很高的可靠性. 同时, 处于不同情境的车辆, 如匀速行驶在稀疏道路上的车辆和马上要经过十字路口的车辆, 对状态更新及时性的要求可能也会有很大的不同, 显然, 后者较前者对及时性的要求更高. 可见, 如果只关注 Level-1 层次的可靠性和延时, 或是在关注可靠性的同时不考虑延时约束, 或是在关注延时的同时不考虑可靠性约束, 问题并不复杂, 传统的自动重传 ARQ (automatic repeat-request)、分集 (diversity) 传输、短包 (short-packet) 通信等技术均会发挥一定的作用. 问题的难点在于 Level-2, Level-3, 甚至是 Level-4 的可靠性与延时保障, 且对可靠性和延时同时进行约束的场景. 更困难的是, 可靠性所要求的是在保障平均延时的基础上还要保证较低的延时抖动, 甚至是确定性延时 (即延时超过某个确定性上界的概率要足够小), 这对业务需求与网络资源均高度动态变化的移动通信网络来讲是一个全新的挑战.

总之, 面向未来多样化的垂直应用, 特别是远程实时监测与控制类应用, 移动通信系统与网络能力建设的重心需要实现以下转变:

(1) 从关注容量向关注可靠性演进;

(2) 从关注通信延时向关注信息时效性演进;

(3) 从关注平均延时向关注确定性时延演进.

3 uRLLC^3 排队论的几点启示

众所周知, 排队论 (queueing theory) 是一门研究随机业务与网络环境下, 为保障可靠性与延时等服务质量要求, 业务需求与网络资源相互匹配的理论, 其中的很多结论对解决上述挑战具有很好的启示作用. 本节将从排队论中几个超越直觉的定理出发, 尝试为 uRLLC^3 的实现提供一些有益启示, 同时给出面向信息时效性的最优调度策略以及面向能量延时折中的最优休眠策略.

造成网络性能下降, 特别是较大延时抖动的原因有很多, 包括网络的拓扑结构、资源配置、路由与流量控制算法、业务量等. 但一般来讲, 在平均意义上网络的设计容量一定会大于需求, 即如果业务需求和资源配置都是确定性的, 那么网络不会发生拥塞. 换言之, 业务与资源的随机性 (randomness) 才是造成网络拥塞的最主要原因, 这就决定了通信网的性能分析需要概率模型以及随机过程理论. 其中最主要的理论之一是排队论 (queueing theory), 它是专门研究业务与网络随机性以及稀缺资源对网络性能影响的基础理论, 起源于 20 世纪初电话网的容量设计 [1], 后被广泛应用于生产系统管理、交通流量控制、互联网容量设计等 [2~4], 其中很多超越直觉的结论可对解决上述挑战提供有益的启示.

启示一: 仅靠提高传输容量并非一定能够实现高可靠低延时, 特别是延时抖动. 传统移动通信网络主要靠扩大传输容量来改善用户的体验, 并成功支撑了移动通信从 1G 到 4G 的演进. 但是显然, 仅靠传统的扩容路线难以满足 5G/6G 网络对 uRLLC^3 的要求.

首先, 扩容并非一定能带来吞吐量和可靠性的提高. 这是因为传输容量 (capacity) 对应的只是传

输能力, 即单位时间内所能传输的最大信息量. 传输容量的提升仅提高了最大可能完成的任务数, 或是降低了完成一件任务的最短时间, 而实际完成的任务数 (即吞吐量) 或在给定时间内成功传输的概率 (即可靠性) 则受限于实际产生的业务需求及其概率分布. 在过去的几十年里, 无线信道一直处在移动互联网端到端通道的瓶颈位置, 因此只要扩容, 吞吐量就会相应地上升. 但随着蜂窝小区的密集化发展, 小区内业务需求的变化越来越剧烈, 小区间业务量的差异也越来越大, 因此扩容不再一定能够带来吞吐量和可靠性的提高. 如果在扩容的同时业务需求没有相应地增加, 那么吞吐量将维持不变. 换言之, 所增加的传输能力被浪费了. 该结论已由贝尔实验室的 Burke 在 1956 年给出了证明 (Burke 定理 [4]), 即无论如何对服务能力进行扩容, 顾客的吞吐量保持恒定.

如果在扩容的同时业务需求也相应地增加的话, 那么系统吞吐量会相应地增加, 但其传输可靠性无法提升. 这可以借用排队论中的爱尔朗 (Erlang) 公式予以说明, 该公式由排队论的创始人 A. K. Erlang 于 1917 年发明, 它给出的是电话业务 (或是任意对延时敏感的实时业务) 在以泊松过程 (Poisson process) 的形式随机到达电话网 (或是任意电路交换网) 时的呼损率, 其形式如下 [4]:

$$P_B = \frac{\frac{a^s}{s!}}{\sum_{j=0}^{s} \frac{a^j}{j!}}, \tag{1}$$

其中, s 表示信道数, $a = \lambda/\mu$ 表示业务量 (λ 表示平均业务到达率, μ 表示平均业务服务率). 可见, 如果在提高容量 (相当于提升 μ) 的同时业务需求也相应升高的话 (相当于 λ 升高, 并保持 a 不变), 则业务呼损率 P_B (可靠性的一种表述) 不会下降. 如果在扩容之后业务需求的随机性有所增加, 即到达过程呈现突发 (bursty) 特性的话, 则业务呼损率反而有可能上升. 只有在扩容之后业务需求及其随机性保持不变或是有所下降时, 可靠性才会提高. 可见, 提高吞吐量和保障可靠性最关键的是传输能力与业务需求的匹配, 一味地扩容并不一定能够提高吞吐量和可靠性.

其次, 扩容也难以降低延时抖动. 所谓延时抖动, 是指实际延时偏离平均延时的相对程度, 一般使用延时的方差系数 ${C^2}_{Wq}$, 即延时的方差与平均值平方的比值来表示. 还是借助于排队论, 考虑一个简单的 $M/M/1$ 队列, 即业务到达过程服从参数为 λ 的泊松过程, 所需服务时间服从参数为 μ 的指数分布, 只有一个服务者, 则业务等待时间 (延时) 的均值与方差系数分别为 [4]

$$\bar{W}_q = \frac{\rho h}{1-\rho}, \tag{2}$$

$${C^2}_{Wq} = \frac{2-\rho}{\rho} > 1, \tag{3}$$

其中, $\rho = \lambda/\mu$ 表示业务负载, $h = 1/\mu$ 表示平均延时. 可见, 如果在提高容量 (相当于提升 μ) 的同时业务需求也相应升高的话 (相当于 λ 升高, 保持 ρ 不变), 则业务延时的平均值会下降, 但延时抖动不会下降, 而且永远大于 1, 表示传输延时会在平均值附近有较大的抖动 (相比于指数分布的抖动). 如果在扩容的同时业务需求并不相应增加, 即 ρ 下降, 则业务延时的平均值会进一步下降, 但延时抖动不仅不会下降, 反而会上升. 换言之, 此时平均延时已经难以反映延时的真实情形. 因此, 虽然扩容可以在一定程度上降低平均延时, 但延时的抖动可能会加大, 使得用户对延时的体验变差.

启示二: 降低业务及资源的随机性是实现高可靠低延时的有效途径. 可见, 为了实现 uRLLC[3], 需要在动态变化的业务需求和网络资源之间建立好匹配机制. 为此, 网络需要变得更加灵活与智能, 即通过人工智能等算法对业务需求做精准预测, 或是通过软件定义或是 AI 驱动等手段实现网络架构与资源配置的动态可调. 同时, 由于 Level-2 级以上的可靠性和延时涉及的不再仅仅是物理层的性能, 而是

包含了排队、调度和计算等过程的全网端到端性能, 且业务需求也会呈现更大的随机性, 因此有必要借助于排队论来寻找灵感.

如前所述, 排队论是研究随机网络环境下业务随机性、受限资源、调度规则等对网络吞吐量、网络延时, 以及传输可靠性等影响的理论. 它起源于电话网的优化设计, 重点关注由于业务随机竞争带来的性能损失, 其核心结论是: 业务随机性是网络性能恶化的主要因素, 且随机性越大, 同等资源下的用户 QoS 越差、资源利用率越低. 因此, 要想提高网络性能, 降低业务及服务过程的随机性是根本.

具体地, 业务到达间隔和服务时间均服从任意概率分布的 $G/G/1$ 排队系统的平均等待时间满足下面的 Kingman 近似公式 [4]:

$$W_q \approx \frac{\rho h({C_a}^2 + {C_b}^2)}{2(1-\rho)}, \tag{4}$$

其中, $\rho = \lambda/\mu$ 为业务负载, $h = 1/\mu$ 为平均服务时间, ${C_a}^2$ 为到达间隔方差系数, ${C_b}^2$ 为服务时间方差系数. 可见, 如果到达间隔和服务时间均不存在随机性, 即等间隔到达和等时长服务 ($ {C_a}^2 = {C_b}^2 = 0$), 则无论业务负载有多大, 平均等待时间均为零. 但如果到达间隔或是服务时间的随机性较大 (${C_a}^2$ 或 ${C_b}^2$ 较大), 即使业务负载较轻, 平均等待时间也有可能很长. 因此, 为了降低延时, 如果业务负载和平均服务时间均难以再降低的话, 比较有效的办法就是降低到达过程和服务过程的随机性.

举例来讲, 针对衰落信道文献 [5] 提出了一种基于注水 (water-filling) 原理的功率控制方案, 其基本概念是在信道条件较好时加大发射功率, 信道条件较差时减少发射功率, 并证明了该方法可使信道容量达到最大. 但这样的功率控制方案不可避免地会造成信道服务率的大幅度波动 (信道条件好时服务率更大, 信道条件不好时服务率更小), 即服务时间的方差系数将变大. 基于上述的 Kingman 近似公式可知, 物理层服务时间的抖动将会加大 MAC 层排队等待的延时, 因此综合来看, 物理层的扩容未必能够带来实际延时的下降.

由此可知, 实现业务需求与网络资源最优匹配的关键是降低业务和资源的随机性. 那么, 在实际系统中, 都有哪些方法可以降低到达过程和服务过程的随机性呢?

(1) 针对业务到达过程, 一个可行的办法是业务汇聚 (traffic aggregation), 因为多个有突发性[1)]业务的叠加 (superposition) 过程将趋于没有突发性的泊松过程 [4]. 当然, 如果原本业务的突发性就很小, 则没有必要进行业务汇聚, 因为汇聚后的随机性反而可能会大于汇聚前的过程. 另一个常用的办法是业务整形 (traffic shaping), 即通过整形器 (shaper) 来调整业务的随机性, 典型的例子是时延敏感网络 (time-sensitive network, TSN) [6].

(2) 针对服务过程, 一个常用的办法是将变化长度的长包拆分成固定长度的短包, 典型的例子就是 ATM (asynchronous transfer mode) 网络, 或是将大块的资源分割成较小的资源块 (resource block, RB), 典型的例子是基于 OFDMA (orthogonal frequency division multiple access) 技术的第四代移动通信系统. 另一个可行的办法是业务卸载 (traffic offloading), 通过分流部分突发业务, 实现业务的均衡. 当然, 还可以在排队规则上做优化, 比如, 如果服务时间是固定的或是抖动较小, 则先到先服务 (first come first serve, FCFS) 的排队模式等待时间抖动最小; 如果服务时间抖动较大, 则给予服务时间较短的业务较高的优先权是最优的, 例如 SJF (shortest-job-first) 策略 [7].

启示三: 提高信息时效性需要联合优化感知、更新与调度策略. 传统的 uRLLC 主要还是面向通信传输过程的优化, 但 uRLLC[3], 即 "信息延时" 及 "信息时效性" 保障还需要综合考虑感知 (sensing)、调度 (scheduling)、决策 (control) 等环节的延时与可靠性[2)], 以及这些性能指标对远程控制系统性能效

1) 这里的 "突发性" 是指方差系数大于 1 的随机过程, 即有时业务集中到达、有时稀疏到达的随机过程.

2) 注意: 无论是感知、调度, 还是决策环节, 都可能需要引入计算 (computing) 过程, 因此这里定义的信息延时包

用的影响. 这在排队论中相当于级联 (tandem) 排队系统, 已有理论 (Jackson 定理 [4]) 告诉我们: 要想最小化端到端的信息延时、保障信息时效性, 需要在感知、调度、决策等环节均衡地部署网络资源, 并对感知、调度、决策等环节的策略进行联合优化.

在感知环节, 为降低信息延时、提高信息时效性, 信源最好是对自身状态或是周边环境连续不断地进行检测, 并立即更新, 即所谓的 "generate-at-will" 工作模式, 使得信宿在有需求时可获得最新的状态信息. 但这势必会消耗信源及网络大量的资源, 尤其是信源的能量资源; 同时也会产生过多的数据包, 从而造成网络的拥塞, 客观上反而会损伤信息时效性. 如果为了节省信源的能量消耗或是减少网络资源的占用而降低检测和更新的频率, 则信宿在有需求时无法获得最新的状态信息, 导致信息延时增大, 损伤信息时效性.

在调度环节, 调度策略不仅需要考虑信源侧状态感知与更新的频率, 而且还要考虑传输资源的约束和不同情境下信源对信息延时的容忍 (敏感) 程度. 一般来讲, 需要优先调度那些状态更新较快, 且对信息延时较敏感的信源. 如果过于频繁地调度某些终端, 且该终端的状态并未发生太大的改变或是并未及时更新, 则可能对改善信息时效性的帮助并不大. 即使该终端的状态发生了较大的变化, 过于频繁的更新不但会消耗终端的能量资源, 还有可能占用更多的网络资源, 损伤其他终端状态信息的时效性.

在决策环节, 信宿需要从多个信源搜集尽量新鲜的状态信息, 然后基于优化或是机器学习等方法计算出最优的决策, 可见其计算能力的配置直接决定了该环节的计算延时. 当然, 如果为了最小化端到端的信息延时而过多地配置计算资源, 则会带来能耗成本的急剧上升. 因此, 需要在不同的情境下针对不同的终端差异化部署计算资源和决策算法.

可见, 为同步提高所有终端状态信息的时效性, 网络中需要尽量减少从终端状态采样到状态信息到达控制中心的端到端延时, 从而使控制中心每次接收到的状态信息更新鲜; 还需要尽可能频繁地更新状态信息, 使控制中心的状态信息保持新鲜; 同时, 针对不同情境, 需要满足差异化的服务质量需求. 在终端能量资源和网络传输资源均受限的情况下, 如何设计终端的感知和更新机制以及网络的调度与服务机制以提高对状态更新业务的服务能力仍需要进一步研究.

4 基于信息年龄的状态更新策略

首先, 为了描述状态更新系统中的信息延时, 文献 [8] 提出了信息年龄 (AoI) 的概念, 它定义为信宿 (如控制中心) 所掌握的最新状态信息距其产生时刻所经历的时间 (或称偏差值), 即

$$h(t) = t - \max(S_i | D_i < t), \tag{5}$$

其中 S_i 为第 i 个成功接收到的数据包产生时间或称时间戳 (timestamp), 而 D_i 为该数据包的接收时间. 如图 2 所示, S_1 时刻采样产生的数据包在 D_1 时刻被成功接收, 则 D_1 的信息年龄从 D_1 下降为 $D_1 - S_1$. 现有研究大多数关注两个性能指标: (1) 平均信息年龄 (average AoI); (2) 峰值信息年龄 (peak AoI). 平均信息年龄指在一段时间内接收到数据包信息年龄的均值, 即图中着色区域面积与总时长之比. 而峰值信息年龄指在信宿接收到更新数据包之前时刻的信息年龄 (图中的尖峰值), 如 D_2 时刻的峰值信息年龄为 $D_2 - S_1$.

可见, AoI 是在信宿 (如控制中心) 端衡量某终端状态信息新鲜度 (information freshness) 的性能指标, 其值越小表示信宿所保有的状态信息越接近于该终端在当前时刻状态信息的真实值, 基于此所

含了计算延时.

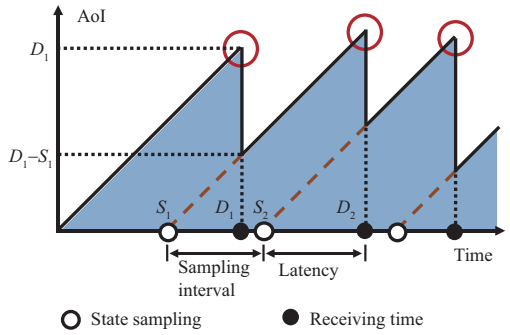

图 2 (网络版彩图) 信息年龄演化曲线

Figure 2 (Color online) Evolution curve of the age-of-information

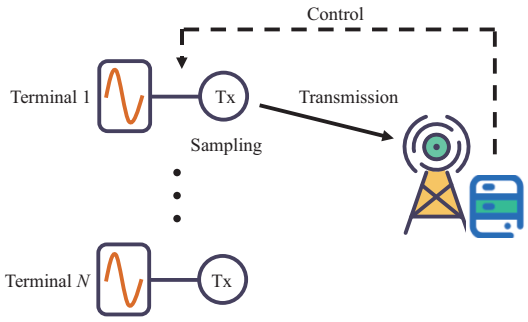

图 3 (网络版彩图) 状态更新系统示意图

Figure 3 (Color online) Schematic diagram of the status update system

作的决策也就有更高的可靠性. 如果为了减少终端感知和更新的能量消耗而降低状态信息感知和更新的频率, 则会导致信息新鲜度变差, 使得信宿难以实时掌握最新的状态; 但如果感知或更新的频率过大, 则终端的能量消耗会急剧上升, 而且有可能由于产生过多的状态更新包造成网络拥塞, 间接地也会导致信息新鲜度变差. 换言之, 取得较好的 AoI 性能既需要信息能够经常被更新 (即较高的可靠性), 也要保证更新得到的信息足够新鲜 (即较低的信息延时), 可见 AoI 可以作为 uRLLC[3] 业务比较理想的性能指标[3)], 亟须研究可使所有终端的平均 AoI 最小的状态更新策略. 未来该研究方向有很大的发展空间, 包括信息新鲜度对信道估计、波束成形与跟踪、功率控制、资源优化等性能的影响.

下面给出一种 AoI 最优的状态信息更新与调度算法 [9]. 考虑 N 个终端向控制中心进行状态更新的离散时间系统 (图 3), 假设状态信息采样过程服从伯努利 (Bernoulli) 分布, 每时隙终端 n 以概率 λ_n 进行状态采样, 每个终端也仅保留最近一次采样数据 (以下简称 "1- 缓冲模型"). 考虑受限的通信资源, 假设每时隙仅支持一个终端上报最新状态信息, 其在控制中心处的信息年龄记作 $h_n(t)$. 令 $d_n(t)$ 表示缓冲中采样数据的等待延时, 即从该数据的产生时刻至当前 t 时刻所经过的时间, 并定义 $u_n(t) \in \{0,1\}$ 为状态更新指示变量, 即 $u_n(t)=1$ 表示终端 n 在时刻 t 进行了状态更新. 由此可知

$$h_n(t+1) = (1-u_n(t))h_n(t) + u_n(t)d_n(t) + 1. \tag{6}$$

则优化加权 (w_n) 长时平均信息年龄的用户调度策略 π 应满足

$$\begin{aligned} \text{P1:}\quad & \min_{\pi} \quad \lim_{T\to\infty} \sup \frac{1}{TN} \sum_{t=1}^{T}\sum_{n=1}^{N} w_n \mathbb{E}_{\pi}[h_n(t)] \\ & \text{s.t.} \quad \sum_{n=1}^{N} u_n(t) \leqslant 1, \quad \forall t \geqslant 1. \end{aligned} \tag{7}$$

引入变量 $r_n(t)$ 表示终端 n 进行状态更新所能带来的信息年龄减小值, 即 $r_n(t) \triangleq h_n(t) - d_n(t)$, 则可将问题 P1 建模为马尔可夫决策过程, 其状态表示为 $S(t) \triangleq \{(r_1(t), d_1(t)), \ldots, (r_N(t), d_N(t))\}$. 如果用户 n 没有被调度, 则系统状态不会更新, 因此信息年龄减小值 $r_n(t)$ 不变, 但采样数据包在队列中的等待延时 $d_n(t)$ 将增加 1 (相当于信息年龄增加 1), 其对应的状态转移概率为

$$P\{(r_n(t), d_n(t)) \to (r_n(t), d_n(t)+1)\} = 1-\lambda_n, \tag{8}$$

3) 由于这里不再仅仅考虑通信过程的可靠性与延时, 因此 uRLLC 中的 "C" 既表示 "Communication", 又表示 "Computing" 和 "Control".

$$P\{(r_n(t),d_n(t))\to(r_n(t)+d_n(t),1)\}=\lambda_n. \tag{9}$$

如果用户 n 被调度, 则系统状态会发生更新, 因此信息年龄将减少至 0, 且采样数据包在队列中的等待延时 $d_n(t)$ 将增加 1, 则其对应的状态转移概率为

$$P\{(r_n(t),d_n(t))\to(0,d_n(t)+1)\}=1-\lambda_n, \tag{10}$$

$$P\{(r_n(t),d_n(t))\to(d_n(t),1)\}=\lambda_n. \tag{11}$$

该马尔可夫决策过程的优化目标为

$$\min_{\pi}\ \lim_{T\to\infty}\sup\frac{1}{TN}\sum_{t=1}^{T}\sum_{n=1}^{N}w_n\mathbb{E}_{\pi}[d_n(t)+r_n(t)(1-u_n(t))]. \tag{12}$$

文献 [10] 已证明该马尔可夫决策过程存在一个具有平稳性质最优调度策略, 但不同用户之间的决策相互影响, 导致系统状态空间随用户数 N 指数增长. 为此, 文献 [9] 引入松弛约束条件 (将约束条件从 "每时隙至多调度一个用户进行更新" 松弛为 "平均每时隙调度用户数不超过 1"), 并引入拉格朗日乘子 β (相当于进行更新所需要付出的额外代价) 对终端决策进行解耦, 得到了单终端子问题, 并证明了解耦后子问题的最优策略具有阈值结构, 即给定状态 (r,d)[4], 当 $r\geqslant R_d$ 时, 最优决策是进行传输; 否则不进行传输, 其阈值 R_d 由下式给出:

$$R_d=\begin{cases}\left|(1-\lambda+d\lambda)J^*-d+1-\lambda\frac{d(d-1)}{2}-\frac{1}{\lambda}\right|, & \text{当 } 1\leqslant d<R_1 \text{ 时},\\ s\left|\lambda\beta\right|, & \text{当 } d\geqslant R_1 \text{ 时},\end{cases} \tag{13}$$

其中长时平均开销 J^* 满足

$$\beta=\left(R_1-1+\frac{1}{\lambda}\right)J^*-\frac{{R_1}^2}{2}+\frac{R_1}{2}-\frac{R_1}{\lambda}+\frac{\lambda-1}{\lambda^2}. \tag{14}$$

基于解耦后子问题的最优策略可设计多用户调度机制. 由于拉格朗日乘子 β 可被视为进行更新的代价, 可见针对子问题求解最优策略的过程相当于计算在给定的信息年龄和采样时延条件下, 进行更新所获得的收益 (即提供更新鲜的状态信息) 是否超过其代价 β. 这与怀特索引 (Whittle index) 概念是很相近的, 即在状态 (r,d) 下的怀特索引值 $I(d,r)$ 等于使得该状态下进行传输更新和不进行传输更新的收益相同的最小 β. 这一点类似于经济学中的均衡思想, $I(d,r)$ 衡量了在此状态下进行更新传输所能获得的收益. 换言之, 该索引反映了用户终端进行状态更新的优先级, 即每时隙永远调度怀特索引值最大的用户终端进行更新传输, 则该策略具有渐近最优性和线性复杂度, 易于实现. 由此, 文献 [9] 给出了怀特索引值的闭式表达式, 如下所述.

定理1 给定采样概率 λ 和状态 (d,r), 记

$$x\triangleq\frac{r+\frac{d(d-1)}{2}\lambda}{1-\lambda+d\lambda}. \tag{15}$$

如果

$$r>\frac{\lambda}{2}d^2+\left(1-\frac{\lambda}{2}\right)d, \tag{16}$$

4) 为了简化符号表示, 以下推导中省略下标 n 及 (t) 表述, 并将常数 w_n 吸收到 β 中.

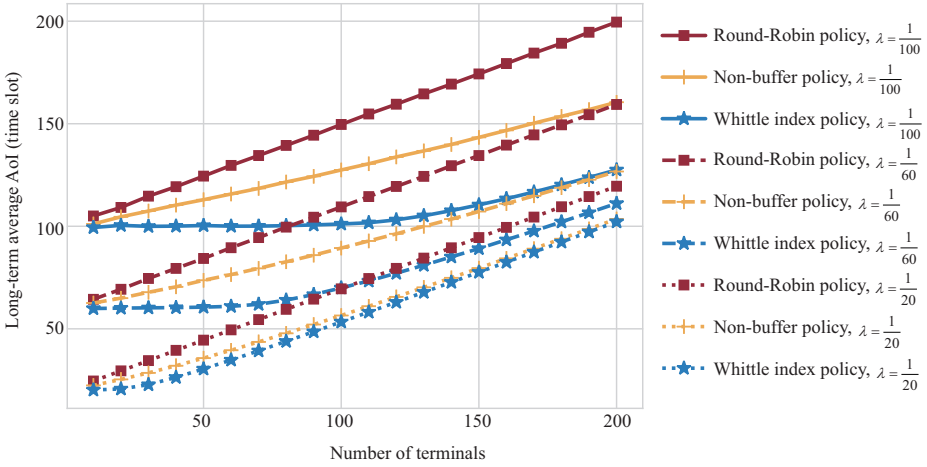

图 4 (网络版彩图) 不同终端数目情况下策略性能比较

Figure 4 (Color online) AoI performance comparison of different policies under different number of terminals

则怀特索引为

$$I(d,r,\lambda)=w\left(\frac{1}{2}x^2+\left(\frac{1}{\lambda}-\frac{1}{2}\right)x\right), \tag{17}$$

否则怀特索引为

$$I(d,r,\lambda)=w\frac{r}{\lambda}. \tag{18}$$

可见, 给定信息年龄, 随着采样延时增加, 怀特索引值先是以二次速率下降, 然后转化为线性速率下降.

图 4 展示了所提策略在较大网络规模下的性能. 在该仿真中, 假设所有用户终端的采样概率相同, 通过调整终端数目得到了不同策略的性能曲线. 图 4 比较了不同采样概率情况下, 轮询 (round-robin, RR) 策略及 0- 缓冲策略与所提策略的性能. 可以看到, 随着终端数目 N 的增长, 三种策略下的平均信息年龄最终均呈现出随着 N 线性增长的趋势. 当采样概率 λ 较大时, 随着终端数目增长, 0- 缓冲策略性能与所提策略相近. 反之, 采样概率较小时, 未在采样时刻进行传输的数据仍然具有较高价值, 因此所提策略性能显著优于 0- 缓冲策略.

文献 [11] 将场景拓展至周期性采样与差错信道情况, 同时给出了各终端当前采样信息等待延时信息不可知情况下的分布式竞争接入信道的最优调度策略, 即各终端仅在其怀特索引大于给定阈值时以一定概率竞争接入信道. 通过该机制可降低随机接入机制下不同终端同时接入带来的冲突概率. 仿真显示, 与已有工作相比, 所提出的集中式调度策略在状态采样率较低时可显著降低平均信息年龄. 同时, 随着数据传输时间增长, 分布式接入与集中式调度策略性能差距逐渐缩小. 鉴于篇幅所限, 本文不再赘述.

进一步地, 文献 [12] 将场景扩展到了吞吐量最优与 AoI 最优异构业务同时共存的情况. 在实际网络中, 状态更新业务需要与通信网络所支撑的其他类型业务共享通信资源, 前者的 QoS 需求一般为最小化信息年龄, 而后者的 QoS 需求一般为带时延约束的高吞吐量. 假设吞吐量最优业务到达率由外界给定, 而状态更新业务采用即时采样模型, 即终端被调度时立即采样获得并上报感知到的信息, 因此问题可描述为: 如何在满足带时延约束的高吞吐量业务服务质量要求的前提下优化状态更新业务时效性. 首先, 基于李雅普诺夫优化理论, 将带时延约束的高吞吐量业务吞吐量约束转化为对应的虚拟队列稳定性要求, 并使用二次形式的信息年龄函数以衡量状态更新业务优先级, 综合虚拟队列与优先级

函数, 设计了最大权重策略. 理论分析给出了迭代估计加权信息年龄下界的算法, 并在此基础上分析了估计误差. 数值仿真显示, 当状态更新业务终端状态采样概率提高时, 所提策略性能与理论极限差距逐渐减小; 而在满足带不同时延约束的高吞吐量业务服务质量要求情况下, 所提策略性能均接近理论极限. 同时, 随着采样率提高, 被调度终端的采样延时逐渐减小. 另外, 仿真也探究了所提策略中业务间权重参数 V 的影响, 发现了时延吞吐量收敛速度与信息年龄性能之间的折中关系, 为实际部署所提策略时的权重参数选择提供了指导.

针对网络中存在大量终端接入与服务的 mMTC 场景, 文献 [13] 给出了一种渐近最优的终端调度与状态更新算法, 有效地解决了网络规模可扩展问题.

5 基于信息紧迫度的状态更新策略

然而, 信息年龄仅表征了状态信息本身的时间属性, 并未反映其对系统性能或效用的影响. 同时, 它也只能表征状态信息新鲜度线性变旧的程度, 无法区别对待不同情境 (context)、不同应用对 AoI 的非线性敏感程度. 实际系统中, 信息新鲜度对系统性能或效用的影响不一定是线性的, 而且不同的情境和不同的应用对 AoI 的敏感程度也可能会有较大差异. 例如在车辆编队驾驶 (platooning) 场景中, 由于车辆移动性较强, 会造成编队所处情境快速变化, 在某些情境下, 例如车辆编队经过路口、车辆编队中有成员离开或是有新成员加入, 或是头车紧急刹车时, 头车的状态信息应当更加频繁地向跟随车辆进行更新, 以保证交通安全. 因此, 受限的无线通信资源应当主要分配给这些情境快速发生变化的车辆. 另一方面, 车辆在离开路口或在稀疏道路上行驶时, 其状态信息新鲜度对周边车辆驾驶安全的影响相对较小, 因此无需过于频繁地更新, 其对资源的占用也可以少一些. 再考虑一个车辆远程控制跟踪应用场景, 车辆在路边单元的控制下按照预定轨迹行进, 在每个时隙开始时, 远程控制器通过已有的状态信息选择某个控制决策, 使得车辆状态尽可能接近预设轨迹. 由于上行信道约束及延时的存在, 车辆的最新状态信息并不是实时可知的, 控制器需要根据历史的状态更新和控制决策对当前时刻的终端状态作出估计, 并根据估计状态进行当前时刻的控制决策. 此时, 跟踪控制的性能一般通过加权平方误差 (weighted mean squared error) 来衡量, 并非信息延时的线性函数, 其中权重体现当前情境下控制精度的重要性.

为此, 我们扩展了信息年龄的概念, 给出了信息紧迫度 [14] (UoI) 的度量指标, 其定义为

$$F(t) = \omega_t \delta(Q_t), \tag{19}$$

其中, ω_t 为状态信息在 t 时刻的情境权重, 若不考虑时变的情境信息, 则 $\omega_t = 1$; Q_t 为信源在 t 时刻的实际状态 x_t 与信宿处所掌握状态 x'_t 之间的偏差值, 两者均随时间变化; $\delta(y)$ 为偏差损失函数, 可根据不同系统和应用定义不同的形式, 例如绝对偏差、二次函数偏差、指数偏差等. 如果选择 $\delta(y) = y$ 且 $\omega_t = 1$, 则 UoI 退化为 AoI. 可见, UoI 是一个普适的信息时效性度量, AoI 是 UoI 的一个特例.

UoI 衡量了由于状态信息更新的不及时对系统性能造成损失的程度, 它不仅刻画了状态信息偏差对实时控制性能的非线性影响, 也反映了状态信息随情境和时间变化的特征. 通过设计状态采样 (感知)、调度和传输策略以减小信息紧迫度, 能够更好地保证状态更新的时效性. 基于这种新的时效性指标在自动驾驶场景的应用, 下面给出了几种 UoI 最优的状态信息调度算法 [14~17], 验证所提 UoI 指标的有效性.

考虑一个离散时间系统, 每个时隙长度为一次状态信息传输所需的时间. 在每个时隙开始前, 状态更新终端根据其当前的情境和状态决定是否进行状态传输. 若终端在 t 时隙进行状态更新, 则记

$U_t=1$, 否则 $U_t=0$. 假设上行信道是一个块衰落信道 (block fading channel), 每个时隙终端的传输成功概率为 p. 记终端在 t 时隙的信道状态为 S_t. 若在 t 时隙进行状态传输能够成功被控制中心接收, 则 $S_t=1$, 否则 $S_t=0$. 因此, 当且仅当 $U_t=S_t=1$ 时状态更新才能成功. 若某次状态更新失败, 即 $U_t=1$ 且 $S_t=0$, 由于假设状态实时可知, 下一次被调度时将仅发送最新的状态信息, 不需要对传输失败的数据包进行重传. 考虑损失函数为二次函数, 则信息紧迫度长时平均的最小化问题描述为

$$\begin{aligned}&\min_{U_t}\ \lim_{T\to\infty}\sup\frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1}\omega_t {Q_t}^2\right]\\&\text{s.t. } Q_{t+1}=(1-U_tS_t)Q_t+A_t,\\&\lim_{T\to\infty}\sup\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[U_t\right]\leqslant\rho,\end{aligned}\tag{20}$$

其中 A_t 为 t 时刻估计误差的增量, ρ 为平均更新频率约束, 反映了终端更新的能耗代价或是传输资源的稀缺性.

由于无线传输的时隙长度一般远小于情境变化的时间尺度, 因此以下首先假设根据当前时刻的情景权重 ω_t 可以预测下一个时隙的情境权重 ω_{t+1}, 并定义在 t 时隙终端的状态更新指数 (update index) J_t 为

$$J_t=\left(\omega_{t+1}-\bar{\omega}+\frac{\bar{\omega}}{p\rho}\right)p{Q_t}^2,\tag{21}$$

其中 $\bar{\omega}$ 为情景权重的长时平均. 可以证明, 相比于不进行状态传输, 进行状态传输后未来所有时隙的信息紧迫度之和的减少量的期望即为状态更新指数. 因此, 状态更新指数表征了在当前时隙进行状态更新对当前和将来的信息紧迫度的降低程度. 由此可给出一种可使信息紧迫度最小的状态更新策略为

$$U_t=\begin{cases}1, & \text{if } J_t>VH_t,\\0, & \text{if } J_t\leqslant VH_t,\end{cases}\tag{22}$$

其中, H_t 为虚拟队列, 满足迭代公式 $H_{t+1}=[H_t-\rho+U_t]^+$; 收敛参数 V 为 Lyapunov 函数中虚拟队列长度 H_t 的权重 [10], 用于调节算法的收敛速度与性能之间的平衡. 因此, 状态更新终端可在每个时隙根据估计误差 Q_t 和下一时隙的权重 ω_{t+1} 计算状态更新指数 J_t, 若其大于 VH_t, 则进行状态更新; 否则选择等待. 这相当于是以状态更新指数来衡量终端的状态更新需求, 并以虚拟队列长度作为动态门限的策略, 当虚拟队列较长, 即前期有过多的状态更新时, 则状态更新的门限值升高, 只有在系统的状态更新指数高过该门限值时才进行传输. 文献 [14] 已证明: 该策略在满足平均更新频率约束下的平均信息紧迫度有上界

$$\lim_{T\to\infty}\sup\frac{1}{T}\sum_{t=0}^{T-1}\omega_tQ_t^2\leqslant\frac{\bar{\omega}\sigma^2}{p\rho}+\frac{1}{2}V.\tag{23}$$

可见, 参数 V 越小, 则信息紧迫度上界越小, 状态更新性能越好. 相反, 参数 V 越大, 虚拟队列长度 H_t 的影响越大, 使得策略对平均更新频率约束更加重视, 同时状态更新策略的收敛速度越快.

图 5 给出了状态更新过程中虚拟队列长度 H_t 和平方误差 Q_t^2 随时间变化趋势. 为了突出在不同情境权重下自适应状态更新策略、虚拟队列长度和平方误差的区别, 在仿真中, 以每 5000 个时隙为一个周期, 在周期中的前 4950 个时隙中情境权重 $\omega_t=1$ (白色背景), 后 50 个时隙的情境权重 $\omega_t=100$ (黄色背景), 即终端有 1% 的时间处于关键情境中, 要求更低的平方误差. 同时, 为了突出状态更新

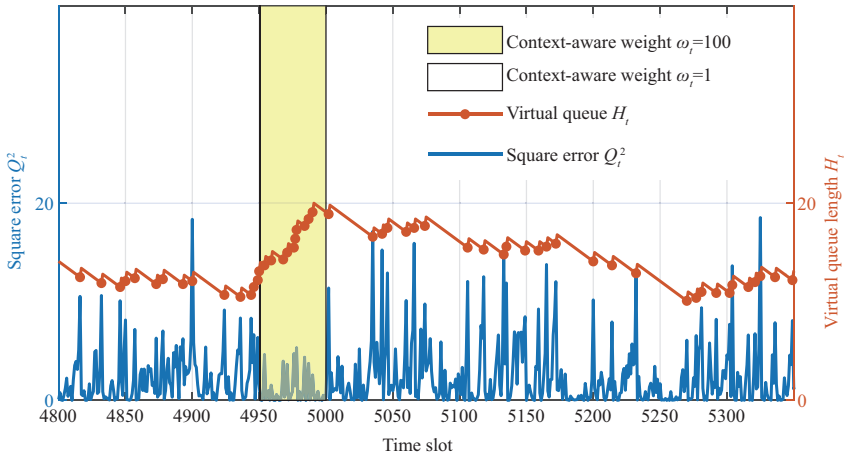

图 5 (网络版彩图) 情景权重发生突变时虚拟队列长度和平方误差随时间变化趋势

Figure 5 (Color online) Variation trend of virtual queue length and squared error over time when the context changes abruptly

策略对状态更新实时性的影响, 令传输成功概率为 $p = 1$, 平均状态更新频率约束 $\rho = 0.1$, 收敛参数 $V = 1$, 误差增量 A_t 为独立同分布的标准高斯随机变量, 即期望为零, 方差为 1. 如图 5 所示, 在关键情境下, 即状态信息重要性较高的阶段, 虚拟队列长度整体上呈上升趋势, 即终端以大于 ρ 的频率进行状态传输, 其状态信息平方误差将明显小于普通情境下的平方误差, 状态更新实时性更高. 在关键情境结束后 (第 5001 时隙之后), 将通过减少传输次数, 仅在平方误差过大时进行状态传输, 以补偿在高情境权重时超过更新频率上限的状态更新所造成的高虚拟队列长度, 使虚拟队列长度减少. 由此可见虚拟队列长度作为动态门限调节状态的效果.

图 6 给出了信息年龄最优策略, 信息紧迫度最优策略, 和自适应状态更新策略在不同平均状态更新频率约束下的信息紧迫度对比. 这里假设传输成功概率 $p = 0.8$, 情景权重 ω_t 以 99% 的概率等于 1, 以 1% 的概率等于 100, 其他参数同图 5. 可见, 尽管在信息紧迫度最优的状态更新策略设计时利用了情境和状态演变的概率分布, 而自适应状态更新策略的设计中只利用了情境权重的均值和误差增量的均值及方差, 在平均状态更新频率上限大于 0.1 时, 自适应状态更新策略下的平均信息紧迫度与信息紧迫度下界几乎重合, 在平均状态更新频率上限小于 0.1 时也非常接近. 同时, 从仿真结果中可以看到, 信息年龄最优的状态更新策略与自适应状态更新策略的信息紧迫度差距较为明显.

上述策略简洁易行, 但需要预测下一时隙的情境权重 ω_{t+1}, 这在实际系统中往往难以实现. 为此, 文献 [15, 16] 提出了一种基于在线强化学习 (reinforcement learning, RL) 算法来学习情境权重变化的方法. 仿真结果显示, 基于 RL 算法的更新策略在更新频率较大时能达到近似最优的性能, 而在更新频率较小时也显著优于基于信息年龄的更新策略. 进一步地, 文献 [17] 考虑了状态更新任务具有计算需求和能量约束情况下的最优任务卸载与更新策略, 并在大规模视频数据集 ILSVRC17-VID 上得到了验证. 对比已有工作及使用额外信息的基线策略, 所提策略的目标追踪成功率更高.

为进一步展示所提状态更新策略的效果, 我们在实验室搭建了车辆编队 (platooning) 驾驶的实物演示平台, 通过车辆间的状态更新实现了多车系统的编队驾驶. 如图 7 所示, 两辆小车跟随头车行驶, 头车不断广播自己的速度状态, 跟随小车接收通信信息并通过激光雷达感知与前车的距离. 若检测到后车位置偏移或后车与前车间距变化, 则对后车的速度和前进方向做相应的补偿. 在无通信时, 由于状态信息更新不及时, 头车紧急刹车时会导致与前车发生碰撞. 在周期性 AoI 更新策略下, 后车可以

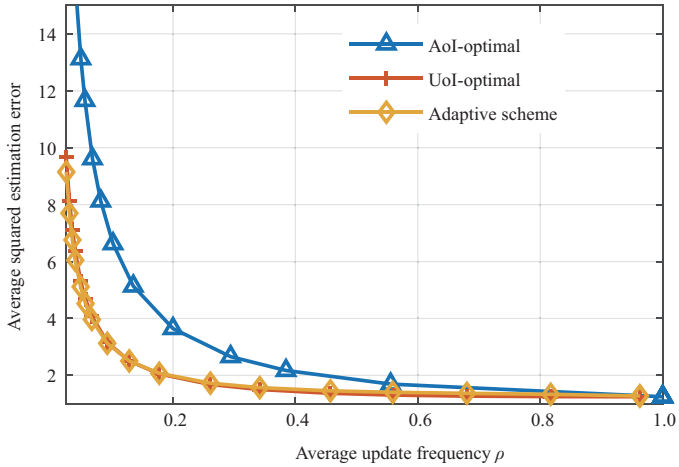

图 6 (网络版彩图) 信息紧迫度最优的更新策略、信息年龄最优的更新策略和自适应状态更新策略性能对比

Figure 6 (Color online) Performance comparison of UoI-optimal strategy, AoI-optimal strategy, and adaptive scheme

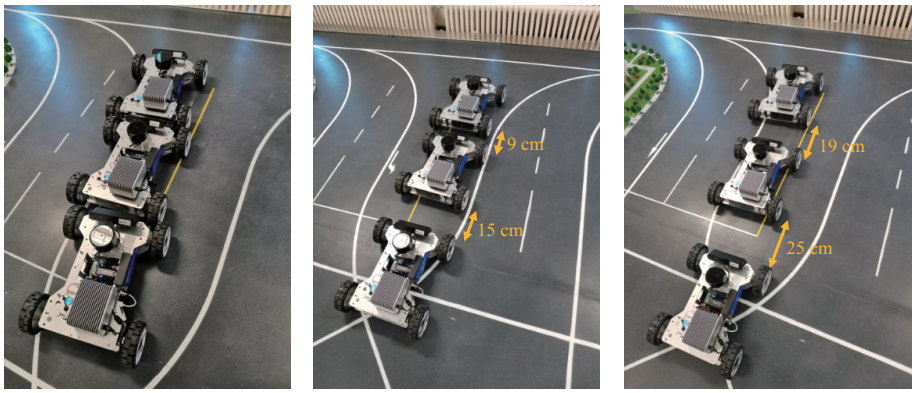

图 7 (网络版彩图) 编队刹车实验图. (a) 车辆之间无通信, 仅依靠激光雷达测量车距进行编队. (b) 车辆之间采用等间隔通信策略, 同时可使用激光雷达实时测量车距. (c) 车辆之间采用基于信息紧迫度的通信策略, 同时可使用激光雷达实时测量车距

Figure 7 (Color online) Brake experiment of vehicle platooning. (a) There is no communication between vehicles, and only LiDAR is used to measure the distance between vehicles for platooning. (b) The vehicles adopt a periodic communication strategy and also use LiDAR to measure the distance between them. (c) The vehicles adopt an UoI-based communication strategy and also use LiDAR to measure the distance between them

及时停下, 但安全距离很小. 在基于信息紧迫度的通信中后车刹车更为及时, 安全距离变大, 其主要原因是车队在匀速行驶过程中, 头车不发送信息, 积攒了通信资源, 因此在头车急刹车的瞬间就可以直接将消息发出, 而周期性更新策略下则需要等到下一个通信时隙. 可见, 自适应状态更新策略能够通过提高状态更新实时性, 有效地减小编队驾驶的跟车距离, 从而提高编队驾驶的效率.

6 总结

本文面向万物智联的 6G 时代, 重点讨论了实现高可靠低延时通信、计算与控制 (uRLLC^3) 的理论与方法, 为实现 6G 网络的超大容量、超高可靠、超低延时, 以及超高能效目标奠定理论与技术基

础. 首先, 针对万物智联应用中的高可靠与低延时内涵进行了扩展, 指出未来 6G 网络需要实现以下 3 个演进: (1) 从关注容量向关注可靠性演进; (2) 从关注通信延时向关注信息时效性演进; 以及 (3) 从关注平均延时向关注确定性时延演进. 然后, 从排队论中几个超越直觉的定理出发, 为 uRLLC^3 的实现提供了几点启示, 即 uRLLC^3 的核心是降低业务模式与资源配置的随机性, 并需要引入新的度量指标. 最后, 以未来智慧网联汽车的智能控制为例, 在描述信息新鲜度的信息年龄 (AoI) 性能指标的基础上, 提出一种新的信息时效性表征指标 —— 信息紧迫度 (UoI), 并在此基础上给出几种 UoI 意义上最优的移动终端调度与控制算法, 验证所提 UoI 指标的有效性.

致谢 本文的完成得到了周盛副教授 (清华大学)、孙宇璇副教授 (北京交通大学)、郑熙博士、孙径舟博士, 以及王乐涵同学的帮助, 在此一并表示感谢.

参考文献

1 Erlang A K. Solution of some problems in the theory of probabilities of significance in automatic telephone exchanges. Post Office Electrical Eng's J, 1917, 10: 189–197
2 Kleinrock L. Queueing Systems: Vol.1: Theory. New York: John Wiley & Sons, 1975
3 Kleinrock L. Queueing Systems: Vol.2: Computer Applications. New York: John Wiley & Sons, 1976
4 Niu Z S, Zhou S. Fundamentals of Communication Networking Theory. Beijing: Tsinghua University Press, 2023 [牛志升, 周盛. 通信网理论基础. 北京: 清华大学出版社, 2023]
5 Goldsmith A J, Varaiya P P. Capacity of fading channels with channel side information. IEEE Trans Inform Theor, 1997, 43: 1986–1992
6 Blokdyk G. Time Sensitive Networking: A Complete Guide. 3rd ed. Brendale: 5STARCooks, 2022
7 Harchol-Balter M. Performance Modeling and Design of Computer Systems: Queueing Theory in Action. Cambridge: Cambridge University Press, 2013
8 Kaul S, Gruteser M, Rai V, et al. Minimizing age of information in vehicular networks. In: Proceedings of the 8th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks, Salt Lake City, 2011. 350–358
9 Sun J, Jiang Z, Krishnamachari B, et al. Closed-form Whittle's index-enabled random access for timely status update. IEEE Trans Commun, 2020, 68: 1538–1551
10 Weber R R, Weiss G. On an index policy for restless bandits. J Appl Probability, 1990, 27: 637–648
11 Sun J, Sun Y, Zhou S, et al. Status update for accurate remote estimation: centralized and decentralized schemes. IEICE Trans Commun, 2022, E105-B: 131–139
12 Sun J, Wang L, Jiang Z, et al. Age-optimal scheduling for heterogeneous traffic with timely throughput constraints. IEEE J Sel Areas Commun, 2021, 39: 1485–1498
13 Jiang Z, Krishnamachari B, Zheng X, et al. Timely status update in wireless uplinks: analytical solutions with asymptotic optimality. IEEE Internet Things J, 2019, 6: 3885–3898
14 Zheng X, Zhou S, Niu Z S. Urgency of information for context-aware timely status updates in remote control systems. IEEE Trans Wireless Commun, 2020, 19: 7237–7250
15 Wang L, Sun J, Zhou S, et al. Timely status update based on urgency of information with statistical context. In: Proceedings of the 32nd International Teletraffic Congress (ITC 32), Osaka, 2020. 90–96
16 Wang L, Sun J, Sun Y, et al. A UoI-optimal policy for timely status updates with resource constraint. Entropy, 2021, 23: 1084
17 Sun J, Wang L, Nan Z S, et al. Optimizing task-specific timeliness with edge-assisted scheduling for status update. IEEE J Sel Areas Inf Theor, 2023, 4: 624–638

uRLLC^3: ultra-reliable and low-latency communication, computing, and control for 6G networks

Zhisheng NIU

Department of Electronic Engineering, Tsinghua University, Beijing 100084*, China*
E-mail: niuzhs@tsinghua.edu.cn

Abstract Ultra-reliable and low-latency communication (uRLLC) is a new service category in 5G and beyond to accommodate mission-critical services and applications, such as autonomous driving and industrial IoT (Internet of Things). But, fundamentally, provisioning of high-reliability and low-latency is contradicting each other, i.e., normally improving reliability may cause longer delay and shortening delay may harm reliability. How to improve both of them is a big challenge, in particular when the service characteristics and available network resources are stochastic and limited in 5G/6G networks. In this regard, as the theory of dealing with stochastic nature of user behavior and resource availability, some perspectives on achieving uRLLC inspired by queueing theory will be discussed. Meanwhile, as 5G/6G and Internet-of-Things (IoT) technologies are deeply integrated into vertical industries such as autonomous driving and industrial robotics, timely status updates (TSU) with ultra-reliable and low-latency are crucial for remote monitoring and controls. In this regard, age of information (AoI) has been proposed to measure the freshness of status updates. However, it is irrelevant to context and just a metric changing linearly with time. We propose a context-aware metric, named as urgency of information (UoI), to measure the nonlinear time-varying importance and the non-uniform context-dependence of the status information. Then we establish a theoretical framework for UoI characterization and provide UoI-optimal status updating and user scheduling schemes. Simulation results show that the proposed updating and scheduling schemes notably outperform the existing ones, such as round robin (RR) and AoI-optimal schemes in terms of UoI, error-bound violation, and control system stability.

Keywords mobile communication, 6G, ultra-reliable and low-latency, vehicular network, age of information, information timeliness